Introduction to SAGEnhaft

Tim Beissbarth

October 28, 2009

1 Overview

Serial Analysis of Gene Expression (SAGE) is a gene expression profiling technique that estimates the abundance of thousands of gene transcripts (mRNAs) from a cell or tissue sample in parallel (Velculescu, 1995). SAGE is based on the sequencing of short sequence tags that are extracted at defined positions of the transcript. As opposed to DNA microarray technology SAGE does not require prior knowledge of the transcripts, and results in an estimate of the absolute abundance of a transcript. However, due to sequencing errors a proportion of the low abundance tags do not represent real genes altering the ability of SAGE to estimate the number of transcripts that have been observed. Moreover, loss of "true"-tags due to sequencing errors will result in altered numbers for the abundance of genuine transcripts.

2 SAGE

Briefly, SAGE works as follows: RNAs from either cells or tissues are converted to double stranded cDNA which is anchored to a solid phase at the 3' end. The double stranded cDNA is then cleaved with a restriction endonuclease at a 4 bp recognition sequence, most commonly CATG. The 3' ends of these cDNA fragments are collected and are then divided into two populations and ligated to linkers containing a type IIS restriction endonuclease recognition sequence, where the enzyme cleaves up to 20 bp away from their recognition site. The two populations are ligated together and amplified by PCR, resulting in two tags orientated tail to tail with an anchoring enzyme recognition site at either end. Two types of SAGE libraries are commonly used, generating tags of different length, i.e. 10 base and 17 base tags respectively, depending on the enzyme used. For protokolls see http://www.sagenet.org/sage_protocol.htm.

3 Base-calling and extraction of SAGE tags

SAGE libraries are generated from between 1,000 to 5,000 sequenced clones, with each sequence run consisting of up to 40 tags. Automated sequencers generate a four-color chromatogram showing the results of the sequencing gel. These chromatograms are read by the Phred or ABI software to call bases and assign an error estimate for each base. These two base-calling programs, the open source program Phred and the ABI KB basecaller, distributed with the ABI 3730 sequencing machines (http://www.appliedbiosystems.com), both assign a quality score to each sequenced base (Ewing 1998). The quality score is given as $-10 \log_{10} P_e$, where P_e is the probability of a base-calling error. The resulting Phred or ABI files are read by functions implemented in this package which subsequently extract the ditags and tags between the anchoring enzyme sites (CATG) in the sequence, keeping the error scores with each base. Ditags have to be within a specified length range, e.g. 20-24 bases for 10 base tags or 32-38 bases for 17 base tags. Duplicate ditags are removed to reduce possible PCR bias, keeping the ditag with the highest average sequencing quality. Tag sequences with a low average sequence quality (≤ 10) are also removed. From experimental SAGE libraries usually 20,000-100,000 tag sequences are generated.

4 Sequence Error correction

Sequencing errors may bias the gene expression measurements made by SAGE. They may introduce non-existent tags at low abundance and decrease the real abundance of other tags. These effects are increased in the longer tags generated in LongSAGE libraries. Current sequencing technology generates quite accurate estimates of sequencing error rates. Here we make use of the sequence neighborhood of SAGE tags and error estimates from the base-calling software to correct for such errors. We introduce a statistical model for the propagation of sequencing errors in SAGE and suggest an Expectation-Maximization (EM) algorithm to correct for them given observed sequences in a library and base-calling error estimates.

For details see: Statistical modeling of sequencing errors in SAGE librarie, **Beissbarth T, Hyde L, Smyth GK, Job C, Boon WM, Tan SS, Scott HS, Speed TP**, Bioinformatics; 7.2004; 20(ISMB Supplement), in press.

5 Comparison of SAGE libraries

SAGE tags are assessed for differential expression between two SAGE libraries by computing Fisher's Exact test for each unique tag. If a particular tag has count n_A in library A and count n_B in library B, and if the total number of sequences counted is t_A for library A and t_B for library B, then Fisher's Exact test is computed to test for independence in the 2×2 contingency table with counts n_A, n_B, $t_A - n_A$ and

$t_B - n_B$. This results in a p-value for the null hypothesis of no differential expression for each gene. Since the tests for different tags are almost independent, the method of Benjamini and Hochberg (1995) was used to control the false discovery rate (fdr). Fisher's Exact test has been found to be slow to compute but an exact binomial test proved to be an excellent approximation when t_A and t_B are large and large relative to n_A and n_B, as they are for typical SAGE libraries. This test is defined similarly to Fisher's Exact test but with binomial probabilities replacing the hypergeometric probabilities. We used a vectorized version of the binomial exact test to allow rapid computation for complete libraries. By analogy with microarray analysis the relative difference of a tag between two libraries is summarized by an M value, which is calculated as $\log_2(n_A + 0.5) + \log_2(t_B - n_B + 0.5) - \log_2(n_B + 0.5) - \log_2(t_A - n_A + 0.5)$, and the mean absolute expression is summarized as an A value, which is calculated as $0.5(\log_2(n_A(t_A + t_B)/2t_A + 0.5) + \log_2(n_B(t_A + t_B)/2t_B + 0.5))$. We call changes with a fdr of less than 0.1 significant.

6 Example

The E15 library was generated from posterior cortex of embryonic C57/BL6 mice at stage E15.5. The B6Hypothal library was generated from hypothalamus of 8 week old C57/BL6 mice.

```
> library(sagenhaft)

> file.copy(system.file("data/E15postHFI.zip",package="sagenhaft"),
+           "E15postHFI.zip")
> E15post<-extract.lib.from.zip("E15postHFI.zip", taglength=10,
+                                 min.ditag.length=20, max.ditag.length=24)

> E15post <- read.sage.library(system.file("data/E15postHFI.sage",
+     package = "sagenhaft"))
> E15post

# libname: E15postHFI
# nseq: 26871
# ntag: 12636
# taglength: 10
# nfiles: 1166
# date: Sun Jun  6 18:27:28 2004
# base.calling.method: seq
# remove.duplicate.ditags: TRUE
# nduplicate.ditags: 231
# remove.N: FALSE
```

```
# remove.low.quality: 10
# min.ditag.length: 20
# max.ditag.length: 24
# cut.site: catg
# EM steps: 50
# likelihood (every 10 steps): 14749.8 15205.7 15220.6 15223.8 15225 15225.5
# var (every 10 steps): 428.5 605.5 610.2 611.1 611.5 611.6
# Removed ShortLinker: 248 249.11
# Removed ShortRibosomal: 982 1042.99
# Removed ShortMitochondrial: 864 881.84
Fields:
libname nseq ntag taglength tags seqs comment
contents of field 'tags':
tag count.raw a c g t avg.ditaglength avg.error.score count.adjusted prop.estimate
contents of field 'seqs':
seq seqextra q1 q2 q3 q4 q5 q6 q7 q8 q9 q10 ditaglength file
```

```
> B6Hypo <- read.sage.library(system.file("data/B6HypothalHFI.sage",
+     package = "sagenhaft"))
> libcomp <- compare.lib.pair(B6Hypo, E15post)
> plot(libcomp)
> libcomp
```

```
# name: B6HypothalHFI:E15postHFI
# ntag: 26589
# taglength: 10
# lib1: B6HypothalHFI
# nseq1: 42775
# ntag1: 18730
# nfiles: 3514
# date: Sun Jun  6 19:08:13 2004
# base.calling.method: seq
# remove.duplicate.ditags: TRUE
# nduplicate.ditags: 214
# remove.N: FALSE
# remove.low.quality: 10
# min.ditag.length: 20
# max.ditag.length: 24
# cut.site: catg
# EM steps: 50
# likelihood (every 10 steps): 36792.6 37547.8 37569.4 37574.1 37575.9 37576.7
# var (every 10 steps): 773.8 1052.4 1059.4 1060.8 1061.3 1061.5
# Removed ShortLinker: 303 303
```

```
# Removed ShortRibosomal: 2795 2911.92
# Removed ShortMitochondrial: 3959 4072.38
# lib2: E15postHFI
# nseq2: 26871
# ntag2: 12636
# nfiles: 1166
# date: Sun Jun  6 18:27:28 2004
# base.calling.method: seq
# remove.duplicate.ditags: TRUE
# nduplicate.ditags: 231
# remove.N: FALSE
# remove.low.quality: 10
# min.ditag.length: 20
# max.ditag.length: 24
# cut.site: catg
# EM steps: 50
# likelihood (every 10 steps): 14749.8 15205.7 15220.6 15223.8 15225 15225.5
# var (every 10 steps): 428.5 605.5 610.2 611.1 611.5 611.6
# Removed ShortLinker: 248 249.11
# Removed ShortRibosomal: 982 1042.99
# Removed ShortMitochondrial: 864 881.84
Fields:
name ntag taglength data comment
contents of field 'data':
tag A.adjusted count1.raw count2.raw M.raw tests.raw count1.adjusted count2.adjusted M.

> testlib <- combine.libs(B6Hypo, E15post)
> testlib <- estimate.errors.mean(testlib)
> testlib <- em.estimate.error.given(testlib)
> tagneighbors <- compute.sequence.neighbors(testlib$seqs[,"seq"], 10,
+                           testlib$seqs[,paste("q", 1:10, sep="")])
```